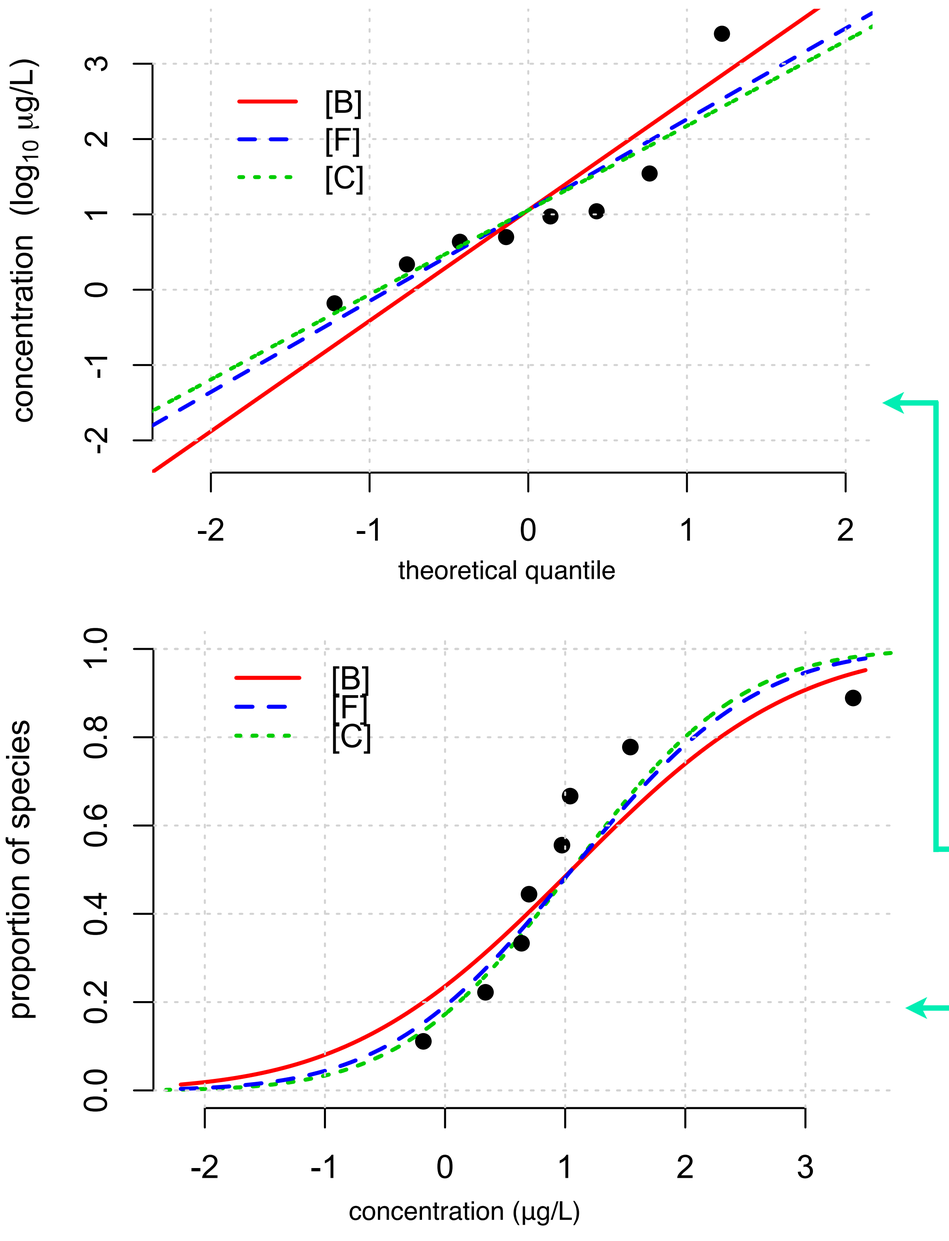
One SSD Model, Three HC_p Estimators: But Which is Better?

Graeme L. Hickey and Peter S. Craig

Durham University, Dept. of Mathematical Sciences, Durham, UK

Email: g.l.hickey@durham.ac.uk

Research kindly funded by Unilever (UK) and Statoil ASA



INTRODUCTION

- The species sensitivity distribution (SSD) is a statistical model of the interspecies variation in tolerance to a toxic stressor within a set of species that may be defined as taxon, assemblage or community.
- The normal distribution (over log transformed concentration) is a standardly used model.
- The hazardous concentration to *p*% of species (HC_p) is defined as the *p*-th percentile of the SSD and is used in regulatory risk characterization and chemical registration.
- Different estimators have been previously used for the (log-)normal SSD by risk assessors. We examine three of them.

FITTED SSDs & HC_p ESTIMATORS

[C] The median pointwise confidence limit (with respect to the sampling distribution of the measured toxicity data) is a regularly used estimator (➔ green dotted line).

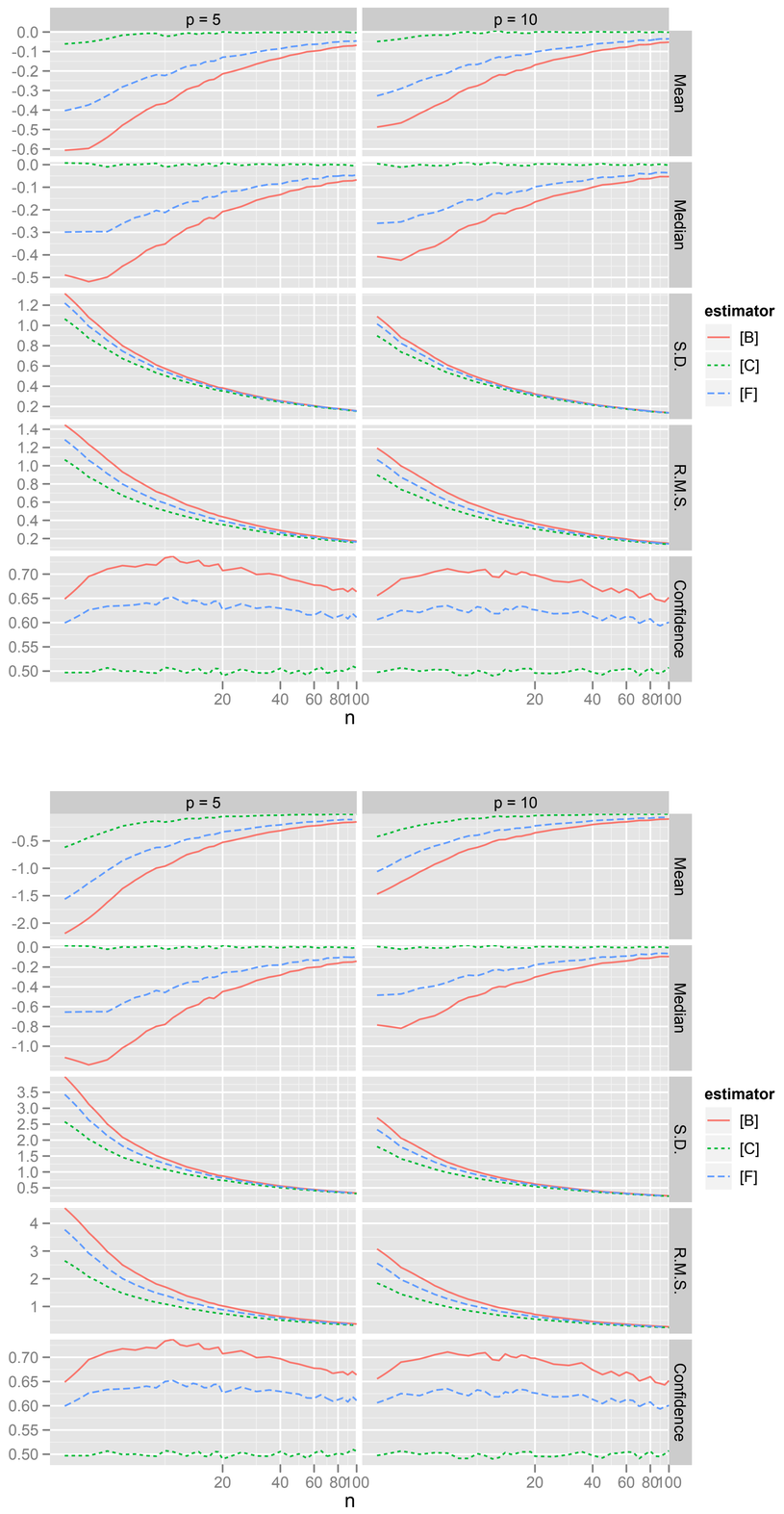
[B] By fitting a linear regression model using ordinary least squares to the numerically ordered log concentration values against the theoretical normal quantiles determined from specified plotting positions (➔ red solid line).

[F] By interchanging the axes of the linear model (i.e. fitting a linear regression to theoretical quantiles against log concentration) a different SSD fit is yielded. This is equivalent to a linear regression by ordinary least squares through a quantile plot (➔ blue dashed line).

Figure 1 (top). Quantile plot of 8 hypothetical log(EC_{50}) values; solid points. The red solid and blue dashed lines are ordinary least square regression line fits: [B] (solid red line) is regression of *X* on *Y*, and [F] (blue dashed line) is regression of *Y* on *X*. The median pointwise confidence limit fit is overlaid (green dotted line); note it is not a straight line. To obtain an estimate of the $\log_{10}(HC_5)$ one would read off the concentration at a theoretical quantile of -1.6445.

Figure 2 (left). A regular 'cumulative distribution function' graphical representation of the 3 SSD fits obtained by transforming the abscissa by the standard normal probability function and interchanging the axes.

- Plotting positions can be chosen in different ways; we used Weibull plotting positions, $q_i = i / (n + 1)$.
- In all cases the $\log_{10}(HC_p)$ estimators are mathematically tractable of the form: $\bar{y} - \kappa(n, p)\hat{\sigma}$ where $\bar{y}$ is the mean of the log transformed toxicity values, $\hat{\sigma}$ is an estimator for the interspecies variance and $\kappa(n, p)$ is a constant which depends on the sample size *n*, protection goal *p* and estimator method.



MEASURING 'OPTIMALITY'

- There is no omnibus method for differentiating between estimators on the grounds of optimality.
- For $p = 5$ and $p = 10$, two measures of discrepancy are analysed:
  - (i) the *standardised* difference between the estimated $\log_{10}(HC_p)$ and the 'true' $\log_{10}(HC_p)$; and
  - (ii) the difference between the log transformed attained fraction of species potentially affected at the estimated HC_p and the intended log transformed protection goal *p*.
- For a range of measured toxicity data sample sizes, the discrepancies were summarized (as shown in Figs. 3 and 4) in terms of the mean (analogous to the statistical bias), median, standard deviation (S.D.), root mean square (R.M.S.; analogous to the mean square error), and confidence (the proportion of estimates which underestimated the intended value).

Figure 3 (left). Summary measures of the standardised discrepancy between the estimated $\log_{10}(HC_p)$ and the 'true' l $\log_{10}(HC_p)$ plotted as functions of sample size *n*. Each curve corresponds to one of the three HC_p estimators.

Figure 4 (bottom). Summary measures of the standardised discrepancy between the log transformed estimated fraction of species potentially affected at the estimated concentration and log(*p*) plotted as functions of sample size *n*. Each curve corresponds to one of the three HC_p estimators.

- Optimal estimators will have a mean and median discrepancy close to zero, small S.D., small R.M.S. and confidence about 0.50.
- The median confidence limit estimator is the closest one in terms of these criteria (NB: its confidence is 0.50 by construction).

CONCLUSIONS

- Regression based HC_p ($p = 5, 10$) estimators are sub-optimal with respect to two discrepancy measures relative to the median confidence limit estimator under the assumption of a log-normal SSD.
- Uncertainty quantification of regression based HC_p estimators is non-trivial because the errors in the linear model are not independent (by virtue of the numerical ordering).
- Regression based HC_p estimators arbitrarily depend of the choice of plotting positions.
- Regression estimators can allow for censored data more easily than the median confidence limit estimator.
- There is no computational benefit in using regression based HC_p estimators since the median confidence limit estimators (as well as other confidence limits) are tabulated in Aldenberg and Jaworska (2000; *Ecotox. Environ. Saf.*, **46**: 1-18) and automated in the RIVM *ETX* 2.0 software program.
- Estimator optimality depends on a chosen loss function (cf. Hickey et al., 2009; *Ecotox. Environ. Saf.*, **72**: 293-300).
- No consistency in the application of estimators by practitioners → source of confusion.

TAKE HOME MESSAGE

For simple SSDs like the (log-)normal distribution with non-censored data, estimate the HC_5 by the median confidence limit and not a regression based estimator.